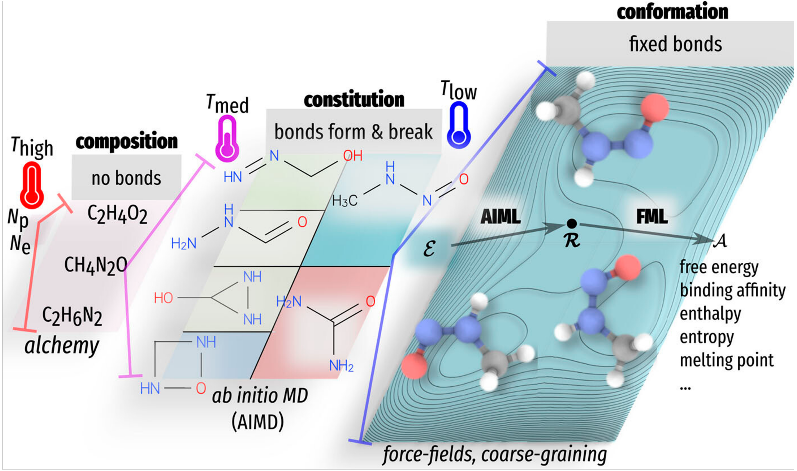
Chemical Space

Hierarchy



https://doi.org/10.1063/5.0095674

Scaling with chemical diversity

Elements	atoms	sum formulas	graphs [1]	conformers [2]
CONF	5	169	4,715	16,797
CONFS	5	349	9,917	51,710
CONFSP	5	757	31,550	
CONFSPCl	5	1,142	37,908	
CONFSPClBr	5	1,647	45,132	
CONFSPClBrI	5	2,291	53,285	328,591

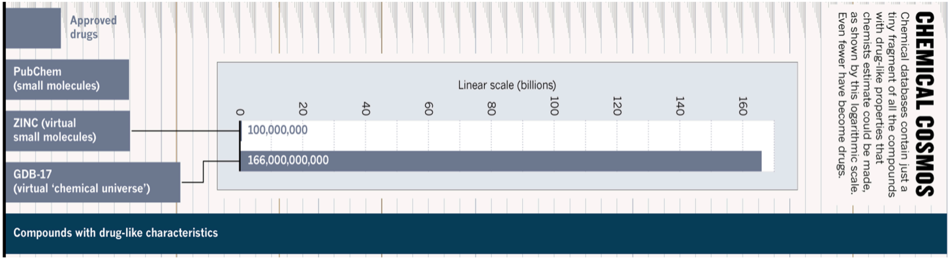
[1] estimated using surge [2] estimated from 32 random graphs with CREST

Scaling with size

Elements	atoms	sum formulas	graphs	conformers[1]
CONF	1	4	4	
CONF	2	19	19	
CONF	3	49	94	
CONF	4	97	621	
CONF	5	169	4,715	
CONF	6	276	42,087	
CONF	7	425	417,923	7,039,390
CONFS	1	5	5	
CONFS	2	28	28	
CONFS	3	82	160	
CONFS	4	180	1,161	
CONFS	5	349	9,917	
CONFS	6	625	97,607	
CONFS	7	1,050	1,064,343	23,016,417

[1] estimated from 32 random graphs with CREST

Chemistry at scale



CHEMICAL COSMOS

Chemical databases contain just a tiny fragment of all the compounds with drug-like properties that chemists estimate could be made, as shown by this logarithmic scale. Even fewer have become drugs.

https://doi.org/10.1038/549445a

Chemistry at scale

Commercial databases

- 164 million molecules
- 15k added daily

Scale

- One person: 1 million compounds/second
- 10 billion people on earth
- 10^{26} universe ages to go through

Necessity vs Convenience

Necessity

- Only way to cover problem size
- Still open to systematic evaluation
- Often used as prefiltering step
- Complicated chemistry
- Tricky / error-prone reference calculations

Convenience

- Can be done more accurately
- Uneconomical/cumbersome reference method
- Often used as direct but optional substitute
- Standard energy calculations of well-behaved systems
- Semi-empirical level sufficient

Sampling Strategies

Random sampling

- Uniform or stratified selection within defined chemical domains
- Useful for statistical benchmarking or coverage estimation
- Inefficient for rare chemistries or constrained systems

Evolutionary sampling

- Guided variation (mutation, crossover) on existing structures
- Mimics chemical evolution or optimization under fitness criteria
- Efficient for property-driven searches but path-dependent
- Methods: Markov-Chain-Monte-Carlo (MCMC) or genetic algorithms (GA)

Design Approaches

Enumeration

- Systematic generation of all combinations up to given limits
- Enables reproducible and exhaustive exploration within constraints
- Exponential scaling; quickly intractable beyond small systems

Generative (rule-based or stochastic)

- Uses chemical rules, reaction networks, or random assembly
- Balances diversity and realism; often less complete than enumeration
- Facilitates targeted coverage of chemically plausible regions

Databases

Computational data

- GDB[1]: molecular graphs (about 166B)
- QM9[2]: small molecules (about 134k)
- QCML[3]: small molecules (33.5M)
- PubChem[4]: from literature
- Mostly energies, rarely other properties

Coverage

- Biased towards conventional molecules
- Synthetically acessible
- Mostly organic

Use

- Benchmarking
- Training data
- Automated method selection

[1] Ruddigkeit et al. *J Chem Inf Model*, 52(11), 2012. [2] Ramakrishnan et al., *Sci Data*, 2014. [3] Ganscha et al., *Sci Data*, 2025. [4] https://pubchem.ncbi.nlm.nih.gov/

Biases in Databases

Selection

- Overrepresentation of small, closed-shell, neutral, organic species
- Expected to be stable
- Neglect of radicals, ions, excited states, and transition structures
- Sampling bias from synthesis feasibility and publication trends

Composition

- Elemental bias (C, H, N, O dominance)
- Energy/geometry bias from relaxed ground-state conformers
- Experimental vs. theoretical data imbalance

Summary

Chemical Space

- Almost none of chemical space has been explored.
- Scaling is a key aspect to think about when comparing methods.
- Chemical diversity drives molecular diversity.